XML Cloud Format Guide

Draft Version 0.4 - September 13, 2018

Purpose

The SRA Sequence Data Delivery Pilot (SDDP) will enable dbGaP users to find and access DNA sequence data which is registered in the NCBI Sequence Read Archive but stored on a supported cloud storage provider. This document provides guidance for using the SRA XML schema to submit metadata for cloud-hosted datasets that will be published through the SRA Sequence Data Delivery Pilot (SDDP). The schema structure is unchanged, but new file types will be supported and certain run attribute key/value pairs are required as part of SDDP submissions. The release and access policies will remain very similar for cloud data submitters. Users with access to patient data will continue to use the approval and access methods managed by dbGaP. The following describes additions to the XML format to support cloud data submissions.

How to Submit Cloud Storage Locations

Only centers submitting XML through a dedicated Aspera account are supported.

Submission of cloud data will use the existing method of submitting a .tar archive containing one <SUBMISSION> XML file with one or both <EXPERIMENT>/<RUN> XML files included. Only the <RUN> XML has been affected by the changes above.

Currently, a single <RUN> record (SRR accession number) can describe either data submitted directly to NCBI or data hosted by others on the cloud, but not both. Therefore cloud data will be submitted through a new run if there is existing data that was previously submitted to NCBI. The new <RUN> can be linked to an existing <EXPERIMENT>, and this would be advised if a cloud location is being added for sequence data currently distributed by NCBI.

Several alternate cloud storage locations for one data file (for example, multiple cloud service providers) can be described in a single <RUN> record. The files, terminal file names and checksums must be identical across storage locations. The example Run XML below shows both Amazon and Google storage locations for a single 1000 Genomes data file.

NCBI will need read access to the cloud bucket for each cloud provider used. Please contact NCBI for details on the necessary bucket permissions.

Example Run XML

ftp://ftp.ncbi.nih.gov/sra/examples/cloud_examples/cloud_run.xml

Temporary XML Schema

ftp://ftp.ncbi.nih.gov/sra/examples/cloud_examples/SRA.cloud.run.xsd

There is a temporary version of the schema available on the NCBI FTP for validating the updated file types. This schema will support all file types SRA expects in cloud submissions. If the official SRA schema is updated to include the cloud submission files, this location will point to the release version of the schema instead.

Run XML Regions Indicated

```xml
<?xml version="1.0" encoding="UTF-8"?>
<RUN xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <IDENTIFIERS>
        <SUBMITTER_ID namespace="submitting_center_abbr">Run_Alias</SUBMITTER_ID>
    </IDENTIFIERS>
    <EXPERIMENT_REF>
        <IDENTIFIERS>
            <SUBMITTER_ID namespace="submitting_center_abbr">Alias_from_submitted_EXPERIMENT</SUBMITTER_ID>
        </IDENTIFIERS>
    </EXPERIMENT_REF>
    <DATA_BLOCK>
        <FILES>
            <FILE filename="NA12878.mapped.ILLUMINA.bwa.CEU.low_coverage.20121211.bam" filetype="bam" checksum_method="MD5" checksum="20e3a973451595736baea13ec91ec3ff"/>
            <FILE filename="NA12878.mapped.ILLUMINA.bwa.CEU.low_coverage.20121211.bam.bai" filetype="bam_index" checksum_method="MD5" checksum="0356980611afcf45b81b4dc5090fea3c"/>
        </FILES>
    </DATA_BLOCK>
    <RUN_ATTRIBUTES>
        <RUN_ATTRIBUTE>
            <TAG>active_location_URL</TAG>
            <VALUE>gs://genomics-public-data/ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/data/NA12878/alignment/</VALUE>
        </RUN_ATTRIBUTE>
        <RUN_ATTRIBUTE>
            <TAG>active_location_URL</TAG>
            <VALUE>s3://1000genomes/phase3/data/NA12878/alignment/</VALUE>
        </RUN_ATTRIBUTE>
        <RUN_ATTRIBUTE>
            <TAG>assembly</TAG>
            <VALUE>GRCh37</VALUE>
        </RUN_ATTRIBUTE>
        <RUN_ATTRIBUTE>
            <TAG>Bases</TAG>
            <VALUE>18280707706</VALUE>
        </RUN_ATTRIBUTE>
        <RUN_ATTRIBUTE>
            <TAG>coverage</TAG>
            <VALUE>5.5</VALUE>
        </RUN_ATTRIBUTE>
        <RUN_ATTRIBUTE>
            <TAG>AvgReadLength</TAG>
            <VALUE>200</VALUE>
        </RUN_ATTRIBUTE>
        <RUN_ATTRIBUTE>
            <TAG>Reads</TAG>
            <VALUE>91014394</VALUE>
        </RUN_ATTRIBUTE>
    </RUN_ATTRIBUTES>
</RUN>
```

Description of Additions to RUN XML format

File types

```
    <FILE filename="NA12878.mapped.ILLUMINA.bwa.CEU.low_coverage.20121211.bam" filetype="bam"
checksum_method="MD5" checksum="20e3a973451595736baea13ec91ec3ff"/>
```

New file types will be supported that are not currently processed by SRA. The new file types include:

Name	Expected Extension	Schema filetype	Description of the Format
Cram Index	.crai	cram_index	Index file for a Cram file. Not useful or expected without a corresponding Cram file.
Bam Index	.bai	bam_index	Index file for a Bam file. Not useful or expected without a corresponding Bam file.
Variant Call Format	.vcf	vcf	Text file with a header containing meta-information followed by data lines describing sequence variation for one or more samples.
Binary Variant Call Format	.bcf	bcf	Binary format of VCF
VCF Index (coordinate)	.csi, .tbi, or .idx	vcf_index	VCF index files including coordinate sorted index for vcf files (.csi), tabix index for vcf files (.tbi), or other formats.

File Count Limit

NCBI will support no more than 10 cloud files in the <FILES> block of the XML of each <RUN> currently. Platforms using a directory tree structure for their data format (Complete Genomics, Oxford Nanopore, etc) will either need to be placed in an archive file or converted to a different format.

Reserved Run Attributes

Storage of the cloud locations will be described by <RUN_ATTRIBUTES> with defined tags.

- **active_location_URL** - Location where data is currently available from. Active locations can't be deleted by XML updates, only marked as inactive by the submitter.

```
<RUN_ATTRIBUTE>
    <TAG>active_location_URL</TAG>
    <VALUE>s3://1000genomes/phase3/data/NA12878/alignment/</VALUE>
</RUN_ATTRIBUTE>
```

  1. Includes the common root to all files in the data block of the run. The combination of (active_location_URL + filename) should provide a valid URL to a single file.
  2. Format is generally <protocol://bucket/prefixes/>. The Amazon cloud protocol is expected to be "s3" while the Google protocol is "gs".
  3. Files can have more than one active location if they are stored on multiple cloud providers or multiple buckets with the same cloud provider.
  4. Only files explicitly listed will be accessible to authorized users through the SRA/SDDP.
  5. NCBI will need read access to the cloud bucket for each cloud provider used. Please contact sra@ncbi.nlm.nih.gov to get details for this.

- **inactive_location_URL** - Location where data existed but is no longer available for download or not present. The submitter can mark an active location as inactive through an XML update.

```
<RUN_ATTRIBUTE>
    <TAG>inactive_location_URL</TAG>
    <VALUE>protocol://bucket/prefixes/</VALUE>
</RUN_ATTRIBUTE>
```

  1. The method for a submitter to mark the loss of access to cloud data
  2. Rules will need to be defined if active files will ever be automatically marked as inactive by NCBI.
  3. Expected to be an update to an existing active_location_URL. The <VALUE> should not be changed but the <TAG> will be changed from <TAG>active_location_URL</TAG> to <TAG>inactive_location_URL</TAG> by the submitter.

Statistics needed for users

- **assembly** - Short name for the assembly used to align the data. Standardized names from GRC or Genome database. (GRCh38, GRCh37, etc)

```
<RUN_ATTRIBUTE>
   <TAG>assembly</TAG>
    <VALUE>GRCh37</VALUE>
</RUN_ATTRIBUTE>
```

- **bases** - Count of unique basecalls present in the data. Please count each base only once if using secondary alignments. For example Samtools summary numbers "SN total length" can be used for this value.

```
<RUN_ATTRIBUTE>
    <TAG>Bases</TAG>
   <VALUE>18280707706</VALUE>
</RUN_ATTRIBUTE>
```

- **reads** - Count of the number of reads in the data. Please count each read only once if using secondary alignments. For example Samtools summary numbers "SN sequences" can be used for this value.

```
<RUN_ATTRIBUTE>
    <TAG>Reads</TAG>
    <VALUE>91014394</VALUE>
</RUN_ATTRIBUTE>
```

- **coverage** - Depth of coverage on assembly used. Found by (Unique Aligned Basecalls)/(Reference Length). For example Samtools summary numbers "bases mapped (cigar)"/(length of the reference assembly used for alignment) can be used for this value.

```
<RUN_ATTRIBUTE>
    <TAG>coverage</TAG>
    <VALUE>5.5</VALUE>
</RUN_ATTRIBUTE>
```

- **AvgReadLength** - Found by (Bases)/(Reads)

```
<RUN_ATTRIBUTE>
    <TAG>AvgReadLength</TAG>
    <VALUE>200</VALUE>
</RUN_ATTRIBUTE>
```